Decision boundary for discrete Bayesian network classifiers

Gherardo Varando GHERARDO.VARANDO@UPM.ES
Concha Bielza MCBIELZA@FI.UPM.ES
Pedro Larrañaga PEDRO.LARRANAGA@FI.UPM.ES
Departamento de Inteligencia Artificial
Universidad Politecnica de Madrid
Campus de Montegancedo, s/n
28660 Boadilla del Monte, Madrid, Spain

Abstract

Bayesian network classifiers are a powerful machine learning tool. In order to evaluate the expressive power of these models, we compute families of polynomials that sign-represent decision functions induced by Bayesian network classifiers. We prove that those families are linear combinations of products of Lagrange basis polynomials. In absence of V-structures in the predictor sub-graph, we are also able to prove that this family of polynomials does indeed characterize the specific classifier considered. We then use this representation to bound the number of decision functions representable by Bayesian network classifiers with a given structure and we compare these bounds to the ones obtained using Vapnik-Chervonenkis dimension.

Keywords: Bayesian networks, supervised classification, decision boundary, polynomial threshold function, Lagrange basis.

1. Introduction

One of the problems with any supervised classification model, and Bayesian network classifiers in particular, is to understand the limits of the expressive power of these models. The first rigorous result in this direction was reported by Minsky (1961), showing that the decision boundary in naive Bayes classifiers with binary predictors is a hyperplane. Since then several other researcher have addressed the problem. Peot (1996) reviewed Minsky's results about binary predictors and presented some extensions. He mainly discussed the case of naive Bayes with k-valued observations and observation-observation dependencies. He also reported an upper bound on the number of linearly separable dichotomies of the vertices of an n-dimensional cube, consequently bounding the number of decision functions that are representable by naive Bayes classifiers with binary predictors. Domingos and Pazzani (1997) studied the optimality of naive Bayes at length and pointed out that, even if the independence assumption among predictors is violated, naive Bayes could achieve optimality under 0-1 loss. Jaeger (2003) showed, for binary predictors that, classifier expressivity at different levels of complexity is characterized by separability with polynomials of different degrees. Ling and Zhang (2003) reported negative results for the expressive power of Bayesian networks; they proved that a Bayesian network where each node has at most k parents cannot represent any function containing $(k+1)$-XORs. Nakamura et al. (2005)

studied the inner product space for Bayesian network classifiers with binary predictors, that is, the smallest Euclidean space that represents the induced concept class. They obtained upper and lower bounds on the dimension of the inner product space and they linked the dimension of the inner product space with the Vapnik-Chervonekis (VC) dimension (Vapnik and Chervonenkis, 1971). Yang and Wu (2012) studied the case of Bayesian networks with k-valued nodes. They computed the VC dimension for fully connected Bayesian networks and for Bayesian networks without V-structures. In both cases they showed that the VC dimension is equal to the dimension of the inner product space.

In this paper we try to generalize the above results within a unified framework. To do this we compute polynomial threshold functions for Bayesian network (BN) binary classifiers in order to express their decision boundaries. This research is restricted to BN classifiers where the binary class variable, C, has no parents and where the predictors are categorical. As usual, our results extend to non-binary classifiers considering an ensemble of binary classifiers. Polynomial threshold functions are a way to describe the decision boundary of a discrete classifier and are a generalization of the results of Minsky (1961) and Peot (1996). In absence of V-structures in the BN we prove that the obtained families of polynomial representing the induced decision functions form linear spaces that are representations of the inner product spaces. We are able to compute the dimensions of those linear spaces and thus of the inner product space extending the results of Nakamura et al. (2005) and Yang and Wu (2012).

In Section 2 we define the notation used and briefly describe Bayesian network classifiers. In Section 3 we define a polynomial representation of the Iverson bracket (Iverson, 1962) over a finite number of categorical variables and derive the representation of discrete probability functions and of conditional probability tables. We then investigate polynomial representations of decision functions induced by Bayesian network classifiers. We look at Bayesian network classifiers in ascending order of complexity: naive Bayes classifiers in Section 3.2, tree augmented naive Bayes classifiers in Section 3.3, Bayesian network-augmented naive Bayes classifiers in Section 3.4 and fully connected Bayesian network classifiers in Section 3.5. In Section 4 we analyse the expressive power of Bayesian network classifiers and we relate our results to the known theory of VC dimension. Finally we present our conclusions and suggest possible future works in Section 5.

2. Preliminaries

We will use bold letters, $\mathbf{x}$ or $\mathbf{k}$, to represent elements of a product space, and letters with a subscript to represent the respective components, e.g. x_2 for the second component of $\mathbf{x}$. The capital letter P always refers to a probability, defined on an appropriate measure space, and capital letters X or X_1, X_2, X_i refer to random variables. For every function $f : \Omega \to \mathbb{R}$ and $\Omega_0 \subseteq \Omega$, we write $f_{|\Omega_0}$ for the restriction of f over Ω_0, that is, the function $f_{|\Omega_0} : \Omega_0 \to \mathbb{R}$ such that $f_{|\Omega_0}(\xi) = f(\xi)$ for every $\xi \in \Omega$.

We consider a binary classification, that is, we are given a training set of labelled observations $\mathcal{T} = \{(\mathbf{x}^1, c^1), \ldots, (\mathbf{x}^N, c^N)\}$, where $\mathbf{x}^i = (x_1^i, x_2^i, \ldots, x_n^i) \in \Omega \subset \mathbb{R}^n$, with $|\Omega| < \infty$, and classes $c^i \in \{-1, +1\}$. We search for a classification algorithm (classifier) Φ that, once trained on the set $\mathcal{T}$, is able to classify every new instance $\mathbf{x} \in \Omega$ into one of the two classes -1 or $+1$. Every classifier induces a decision function $f_{\mathcal{T}}^{\Phi} : \Omega \to \{-1, +1\}$, where the clas-

sifier Φ will classify each new instance $\mathbf{x}$ to class a if $f_{\mathcal{T}}^{\Phi}(\mathbf{x}) = a$. We drop the subscript $\mathcal{T}$ since we are not interested in the relationship to the training set.

In this paper we focus on Bayes classifiers, probabilistic classifiers which learn from the training set $\mathcal{T}$ a joint probability $P(\mathbf{X}, C)$ and classify each new instance $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ in the most probable a posteriori class (MAP), that is,

$$f^{\Phi}(\mathbf{x}) = arg\max_{c} P(C = c|\mathbf{X} = \mathbf{x}) = arg\max_{c} P(\mathbf{X} = \mathbf{x}, C = c).$$

BN classifiers (Bielza and Larrañaga, 2014) are Bayesian classifiers that factorize the joint probability distribution according to a Bayesian network. They range from the simplest naive Bayes classifier (Figure 1), where the predictor variables are assumed to be conditionally independent given the class variable, to the unrestricted Bayesian classifier, where a general form of Bayesian network (Pearl, 1988) is permitted. We will study only Bayesian network augmented naive Bayes classifiers, that is, we will consider the class C as a root node parent of every predictor variable. Once the structure of the Bayesian network is fixed, we need to estimate the parameters of the probability distribution. Thanks to the factorization implied by the Bayesian network structure we just estimate the conditional probability distributions of every variable given its parents, that is we have to estimate $P(X_i = x_i|\mathbf{X}_{\mathbf{pa}(i)} = \mathbf{x}_{\mathbf{pa}(i)})$, where $\mathbf{X}_{\mathbf{pa}(i)}$ stands for the vector of the parents of X_i. In the discrete case this is reduced to the estimation of conditional probability tables. They could be estimated in several ways, but the straightforward approach using the maximum likelihood estimators (MLE), which are the relative frequencies, could lead to some conditional probabilities equal to zero. A Bayesian approach, such as the Laplace estimator or more generally Dirichlet-prior estimation of the parameters, will avoid this drawback. Because of this observation we will assume from now on that all parameters learned will be different from zero, that is, all the probabilities are positive.

To describe the complexity of decision functions we use the concept of threshold functions.

Definition 1 *Given a decision function $f : \Omega \to \{-1, +1\}$, where $\Omega \subset \mathbb{R}^n$, $|\Omega| < \infty$ and $r : \mathbb{R}^n \mapsto \mathbb{R}$ a polynomial we say that r sign-represents f or that f is computed by a polynomial threshold function, if*

$$f(\mathbf{x}) = sgn(r(\mathbf{x})) \text{ for every } \mathbf{x} \in \Omega.$$

Moreover, given a set of polynomials $\mathcal{P}$, we denote by $sgn(\mathcal{P})$ the set of decision functions that are sign-representable by polynomials in $\mathcal{P}$ and by $\{-1, +1\}^{\Omega}$ the set of all the $2^{|\Omega|}$ decision functions over Ω. Polynomial threshold functions are mainly studied in the theory of Boolean functions, functions $g : \{-1, +1\}^n \to \{-1, +1\}$ (O'Donnell and Servedio, 2010; Wang and Williams, 1991). A particular case is the linear threshold function, that is, when the degree of the polynomial that sign-represents the decision function is equal to one. Observe that different polynomials can sign-represent the same decision function, and not every polynomial sign-represents a decision function. In general we have that a polynomial $r(\mathbf{x})$ sign-represents a decision function over Ω if and only if $r(\mathbf{x}) \neq 0$ for every $\mathbf{x} \in \Omega$.

3. Polynomial Threshold Functions for Bayesian Network Classifiers

We develop a method to easily compute polynomial threshold functions for Bayesian network classifiers. This method is an extension of the well-known results on the decision boundary of naive Bayes classifiers (Minsky, 1961; Peot, 1996). The method is based on the polynomial interpolation of discrete probability functions or equivalently their logarithms. Pistone et al. (2001) give a more formal and general description of this subject, also addressing applications to Bayesian networks. We will develop this method directly using Lagrange basis polynomials.

3.1 Lagrange Interpolation of Discrete Probability

The proofs of the results on the decision boundary in naive Bayes classifiers are based on a representation of the categorical distribution over two values $\{0,1\}$ in an exponential form, $P(X=x) = p^x(1-p)^{1-x}$, with $x \in \{0,1\}$ and $p \in (0,1)$. We aim to reproduce the same representation for a categorical variable $X \in \Lambda = \{\xi^1, \xi^2, \ldots, \xi^m\} \subset \mathbb{R}$. We consider $\{p(1), \ldots, p(m)\}$ such that $\sum_{j=1}^m p(j) = 1$ and, using the Iverson bracket (Iverson, 1962), we write

$$P(X=x) = \prod_{j=1}^{m} p(j)^{[x=\xi^j]}. \tag{1}$$

If $X \in \{0,1\}$ we could represent $[x=0]$ as x and $[x=1]$ as $1-x$. When $X \in \Lambda = \{\xi^1, \xi^2, \ldots, \xi^m\} \subset \mathbb{R}$, we need to find m polynomials $\left\{\ell_j^\Lambda\right\}_{j=1}^m$ such that

$$\ell_j^\Lambda(\xi^j) = 1,$$

and

$$\ell_j^\Lambda(\xi^k) = 0 \text{ for every } k \neq j.$$

We easily see that such polynomials exist and have the following form:

$$\ell_j^\Lambda(x) = \prod_{k \neq j} \frac{(x - \xi^k)}{(\xi^j - \xi^k)}. \tag{2}$$

The polynomials defined in Equation (2) are the Lagrange basis polynomials over the points in Λ. These polynomials are m linearly independent polynomials of degree $m-1$, and so they form a basis of polynomials in one variable whose degree is at most $m-1$. We summarize some properties of these polynomials in the following lemma.

Lemma 2 *Let $\Omega_i = \{\xi_i^1, \xi_i^2, \ldots, \xi_i^{m_i}\} \subset \mathbb{R}$, for $i = 1, 2, \ldots, n$. For every i define the Lagrange basis, $\left\{\ell_j^{\Omega_i}(x_i)\right\}$, over Ω_i as in Equation (2). Then we have*

1. *For every $i \in \{1, 2, \ldots, n\}$, $\left\{\ell_j^{\Omega_i}(x_i)\right\}_{j=1}^{m_i}$ form a basis of the space of polynomials in x_i of degree $|\Omega_i| - 1$.*
2. *$\sum_{i \in I} \sum_{j_i=1}^{m_i} \prod_{i \in I} \ell_{j_i}^{\Omega_i}(x_i) = 1$, for every $x_i \in \mathbb{R}$ and every $I \subseteq \{1, 2, \ldots, n\}$.*

3. $\prod_{i\in I} \ell_{j_i}^{\Omega_i}(x_i) = [x_i = \xi_i^{j_i}\ \forall i \in I]$, *for every* $\xi_i^{j_i} \in \Omega_i$ *and* $I \subseteq \{1, 2, \ldots, n\}$.

4. $\sum_{i\in J} \sum_{j_i=1}^{m_i} \prod_{i\in I} \ell_{j_i}^{\Omega_i}(x_i) = \prod_{i\in I\setminus J} \ell_{j_i}^{\Omega_i}(x_i)$, *for every* $x_i \in \mathbb{R}$ *and* $J \subset I \subseteq \{1, \ldots, n\}$.

Proof The proof of the above lemma is trivial, and we merely outline some points. Point 1 follows from the linear independences of the Lagrange basis polynomials. To prove point 2, we have merely to observe that, since $\left\{\ell_j^{\Omega_i}\right\}_{j=1}^{m_i}$ is a basis we have that the polynomial constant 1 admits a unique representation in the considered basis, in particular $1 = \sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i)$. Point 3 follows trivially by substitution, and point 4 is derived from point 2. ■

If we are given a categorical random variable X over $\Lambda = \{\xi^1, \ldots, \xi^m\}$ whose probability mass function is P, we are able to rewrite Equation (1) using the Lagrange basis, as

$$P(X = x) = \prod_{j=1}^{m} p(j)^{[x=\xi^j]} = \prod_{j=1}^{m} p(j)^{\ell_j^{\Lambda}(x)}, \tag{3}$$

where $p(j) = P(X = \xi^j)$ are the values of the probability mass function over Λ. Equation (3) is a consequence of the identity $[x = \xi^j] = \ell_j^{\Lambda}(x)$ which derives from point 3 of Lemma 2 considering $|I| = 1$. More generally, we consider a set of random variables $\{X_1, X_2, \ldots, X_n\}$ such that, for every $i = 1, \ldots, n$, the variable $X_i \in \Omega_i = \{\xi_i^1, \xi_i^2, \ldots, \xi_i^{m_i}\}$. If we are given a conditional probability table that represents the probability function $P(X_1 = x_1 | X_2 = x_2, \ldots, X_n = x_n)$, we can use the Iverson bracket over n variables $x_1, \ldots, x_n$ to describe the conditional distribution of X_1 given $X_2, \ldots, X_n$,

$$P\left(X_1 = x_1 | X_2 = x_2, \ldots, X_n = x_n\right) = \prod_{(j_1,\ldots,j_n)} p(j_1 | j_2, \ldots, j_n)^{[x_i=\xi_i^{j_i}\ \forall i=1,\ldots,n]},$$

where $p(j_1 | j_2, \ldots, j_n) = P(X_1 = \xi_1^{j_1} | X_2 = \xi_2^{j_2}, \ldots, X_n = \xi_n^{j_n})$ are the values of the conditional probability table. Now using point 3 of Lemma 2 with $I = \{1, \ldots, n\}$, we get

$$P\left(X_1 = x_1 | X_2 = x_2, \ldots, X_n = x_n\right) = \prod_{(j_1,\ldots,j_n)} p(j_1 | j_2, \ldots, j_n)^{\prod_{i=1}^{m} \ell_{j_i}^{\Omega_i}(x_i)}. \tag{4}$$

3.2 Naive Bayes

We consider a naive Bayes classifier (NB) (Figure 1) where the predictor variables $X_i \in \Omega_i$ are conditionally independent given the class variable C. The joint probability distribution factorizes as follows:

$$P(C = c, X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(C = c) \prod_{i=1}^{n} P(X_i = x_i | C = c). \tag{5}$$

If the predictor variables are binary, Minsky (1961) proved that the decision boundaries are hyperplanes. For categorical predictors, the scenario is much more complicated as shown in Figure 2.

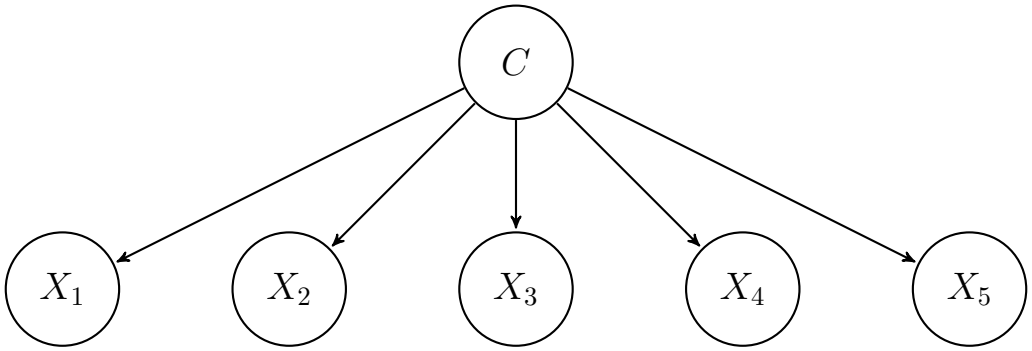
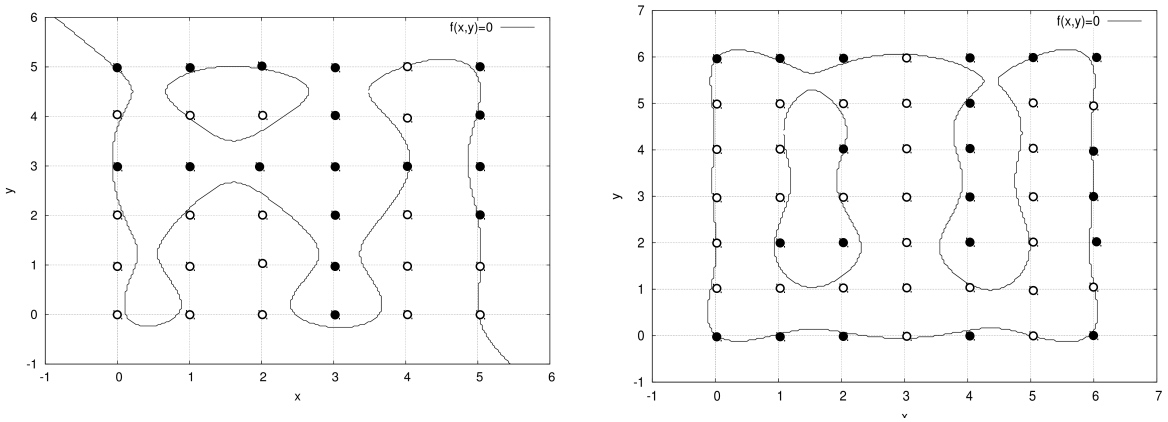


Figure 1: Naive Bayes classifier structure with five predictor variables



(a) $X, Y \in \{0, 1, \ldots, 5\}$ (b) $X, Y \in \{0, 1, \ldots, 6\}$

Figure 2: Decision boundary for two naive Bayes classifiers with two categorical variables X, Y.

Theorem 3 *A decision function f over n categorical variables $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$, with $|\Omega_i| = m_i$, is sign-represented by a polynomial of the form $\sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) \right)$ if and only if there exists a naive Bayes classifier that induces f, where $\ell_j^{\Omega_i}$ are the Lagrange bases over Ω_i.*

Proof We consider a naive Bayes classifier as in Figure 1. For every $i = 1, \ldots, n$ the variable X_i takes values over $\Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$, a subset of $\mathbb{R}$ of cardinality m_i. Thanks to Equation (3), we can express, for every value c of the class, the conditional probability $P(X_i|C)$ as

$$P(X_i = x_i | C = c) = \prod_{j=1}^{m_i} p_i(j|c)^{\ell_j^{\Omega_i}(x_i)},$$

where $p_i(j|c) = P(X_i = \xi_i^j | C = c)$. If we define $a_i(j|c) = \ln(p_i(j|c))$, and assuming that $p_i(j|c) > 0$, we have that

$$P(X_i = x_i | C = c) = \exp\left(\sum_{j=1}^{m_i} a_i(j|c)\ell_j^{\Omega_i}(x_i)\right). \tag{6}$$

Using this representation we easily find the decision function for NB with arbitrary discrete predictor variables. Setting $a = \ln(P(C = +1))$ and $b = \ln(P(C = -1))$, we have that a new instance $\mathbf{x} = (x_1, \ldots, x_n)$ will be classified as $C = +1$ if

$$P(X_1 = x_1, \ldots, X_n = x_n, C = +1) > P(X_1 = x_1, \ldots, X_n = x_n, C = -1).$$

Using Equations (5) and (6) we have that the previous inequality could be rewritten as

$$\exp\left(a + \sum_{i=1}^{n}\left(\sum_{j=1}^{m_i} a_i(j|+1)\ell_j^{\Omega_i}(x_i)\right)\right) > \exp\left(b + \sum_{i=1}^{n}\left(\sum_{j=1}^{m_i} a_i(j|-1)\ell_j^{\Omega_i}(x_i)\right)\right),$$

so the decision function for a naive Bayes classifier is

$$f^{NB}(\mathbf{x}) = sgn\left(a - b + \sum_{i=1}^{n}\left(\sum_{j=1}^{m_i} \alpha_i(j)\ell_j^{\Omega_i}(x_i)\right)\right), \tag{7}$$

where $\alpha_i(j) = a_i(j|+1) - a_i(j|-1) = \ln\left(\frac{P(X_i=j|C=+1)}{P(X_i=j|C=-1)}\right)$. We see from Equation (7) that the decision function is sign-represented by a polynomial function of degree $m-1$ and also the polynomial admits the representation $\sum_{i=1}^{n}\left(\sum_{j=1}^{m_i}\alpha_i(j)\ell_j^{\Omega_i}(x_i)\right)$. We have proved the *if* part of the theorem.

To prove the *only if* we have to find a naive Bayes model that induces the decision function

$$f(\mathbf{x}) = sgn\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{m_i}\alpha_i(j)\ell_j^{\Omega_i}(x_i)\right)\right),$$

for fixed, arbitrary coefficients $\alpha_i(j)$. A naive Bayes model is defined by the probability $P(X_i = \xi_i^j | C = c) = p_i(j|c)$ for $i = 1, \ldots, n$; $\xi_i^j \in \Omega_i$ and $c \in \{-1, +1\}$ subject to the $2n$ constraints $\sum_{j=1}^{m_i} p_i(j|c) = 1$ and by the prior probability over the class $1 - P(C = -1) = P(C = +1) \in (0, 1)$. Due to the fact that $\sum_{j=1}^{m_i}\ell_j^{\Omega_i}(x_i) = 1$ (Lemma 2, point 2 with $|I| = 1$), we can write

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{m_i}\alpha_i(j)\ell_j^{\Omega_i}(x_i)\right) = \sum_{i=1}^{n}\left(\sum_{j=1}^{m_i}\alpha_i(j)\ell_j^{\Omega_i}(x_i)\right) - \alpha + \alpha$$

$$= \sum_{i=1}^{n}\left(\sum_{j=1}^{m_i}\alpha_i(j)\ell_j^{\Omega_i}(x_i)\right) - \sum_{i=1}^{n}\left(\alpha_i\sum_{j=1}^{m_i}\ell_j^{\Omega_i}(x_i)\right) + \alpha = \sum_{i=1}^{n}\left(\sum_{j=1}^{m_i}(\alpha_i(j) - \alpha_i)\ell_j^{\Omega_i}(x_i)\right) + \alpha,$$

where $\sum_{i=1}^{n} \alpha_i = \alpha$. Now the decision function is in the form of Equation (7). We have to find for every $i = 1, \ldots, n$ $\{p_i(j|c)\}_{j=1}^{m_i} \in (0,1)$ such that $\sum_{j=1}^{m_i} p_i(j|c) = 1$ for every $c \in \{-1, +1\}$ and $\alpha_i(j) - \alpha_i = \ln(\frac{p_i(j|+1)}{p_i(j|-1)})$. We expand the above equation to

$$p_i(j|+1) = e^{(\alpha_i(j)-\alpha_i)} p_i(j|-1).$$

We can now choose an arbitrary distribution $p_i(j|-1)$, for example $p_i(j|-1) = \frac{1}{m_i}$ for every $j = 1, \ldots, m_i$, and from the above relation we have that $p_i(j|+1) = e^{(\alpha_i(j)-\alpha_i)} p_i(j|-1)$. We now have to check that these probability assignments satisfy the constraints. Because of the previous choice, the numbers $p_i(j|-1) \in (0,1)$ form a probability distribution over Ω_i so obviously $\sum_{j=1}^{m_i} p_i(j|-1) = 1$. For $p_i(j|+1)$ we have that

$$\sum_{j=1}^{m_i} p_i(j|+1) = \sum_{j=1}^{m_i} p_i(j|-1) e^{\alpha_i(j)} e^{-\alpha_i} = e^{-\alpha_i} \sum_{j=1}^{m_i} p_i(j|-1) e^{\alpha_i(j)}.$$

So we just have to choose α_i, which is still a free parameter, such that

$$e^{\alpha_i} = \sum_{j=1}^{m_i} p_i(j|-1) e^{\alpha_i(j)}$$

in order to get $\sum_{j=1}^{m_i} p_i(j|+1) = 1$. Lastly, to obtain the target naive Bayes model, we just need to set $P(C = +1) \in (0,1)$ such that

$$\alpha = \sum_{i=1}^{n} \alpha_i = \ln\left(\frac{P(C=+1)}{P(C=-1)}\right) = \ln\left(\frac{P(C=+1)}{1 - P(C=+1)}\right).$$

This is possible because the function $\ln\left(\frac{p}{1-p}\right)$ maps $(0,1)$ into $(0,\infty)$. ■

As a result of Theorem 3 we have that a naive Bayes classifier could represent every decision function which is sign-representable by a polynomial of the family

$$\left\{ r(\mathbf{x}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) \right), \alpha_i(j) \in \mathbb{R} \right\}.$$

Only if we fix the prior probability over the class C are there restrictions on the coefficients $\alpha_i(j)$.

Theorem 4 *Let f be a decision function for a binary classification problem with n categorical predictor variables $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\} \subset \mathbb{R}$. The following sentences are equivalent:*

i) f is sign-represented by a polynomial of the form $\sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) \right)$ with $\alpha_i(j)$ such that for every $i = 1, \ldots, n$, there exists $j_{i,1}$ and $j_{i,2}$ such that $e^{\alpha_i(j_{i,1})} < 1$ and $e^{\alpha_i(j_{i,2})} > 1$ or alternatively $e^{\alpha_i(j)} = 1$ for every $j = 1, \ldots, m_i$.

ii) There exists a naive Bayes classifier that induces f, with uniform prior probability over the class C.

Proof To prove that *i)* implies *ii)* we have to find a naive Bayes classifier that induces f, that is, we have to find $P(X_i = j|C = c) = p_i(j|c)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$ such that $p_i(j|c)$ are probability mass functions over Ω_i for every fixed i and $c \in \{-1, +1\}$. Moreover, analogously to the proof of Theorem 3, we require that $p_i(j|+1) = e^{\alpha_i(j)} p_i(j|-1)$. This is equivalent solving the following linear-constrained problem for every $i = 1, \ldots, n$:

$$\begin{cases} \sum_{j=1}^{m_i} p_{i,j} = 1 \\ \sum_{j=1}^{m_i} A_{i,j} p_{i,j} = 1 \\ p_{i,j} > 0 \text{ for every } j = 1, \ldots, m_i, \end{cases} \tag{8}$$

where $A_{i,j} = e^{\alpha_i(j)}$ and $p_{i,j} = p_i(j|-1)$.

From hypothesis *i)* we have two possible cases for every fixed $i = 1, \ldots, n$. If $e^{\alpha_i(j)} = A_{i,j} = 1$ for every j then the problem represented by Equations (8) admits all the $\{p_{i,j}\}_{j=1}^{m_i}$ solutions that are probability mass functions over Ω_i. Otherwise there are $j_{i,1}$ and $j_{i,2}$ such that $A_{i,j_{i,1}} = e^{\alpha_i(j_{i,1})} < 1$ and $A_{i,j_{i,2}} = e^{\alpha_i(j_{i,2})} > 1$; we solve the underdeterminate linear system in Equations (8) with respect to $p_{i,j_{i,1}}$ and $p_{i,j_{i,2}}$, obtaining

$$\begin{cases} p_{i,j_{i,1}} = 1 - \sum_{j \neq j_{i,1}} p_{i,j} \\ p_{i,j_{i,2}} = \frac{1 - A_{i,j_{i,1}}}{A_{i,j_{i,2}} - A_{i,j_{i,1}}} + \frac{\sum_{j \neq j_{i,1}, j_{i,2}} \left((A_{i,j_{i,1}} - A_{i,j}) p_{i,j} \right)}{A_{i,j_{i,2}} - A_{i,j_{i,1}}} \end{cases}$$

We now have to choose the *free parameters* $\{p_{i,j}\}_{j \neq j_{i,1}, j_{i,2}}$ in such a way that $p_{i,j} \in (0, 1)$ for every $j = 1, \ldots, m_i$. This is possible because

$$\lim_{\{p_{i,j}\} \to 0^+} p_{i,j_{i,2}} = \frac{1 - A_{i,j_{i,1}}}{A_{i,j_{i,2}} - A_{i,j_{i,1}}} \in (0, 1),$$
$$\lim_{\{p_{i,j}\} \to 0^+} p_{i,j_{i,1}} = 1 - \frac{1 - A_{i,j_{i,1}}}{A_{i,j_{i,2}} - A_{i,j_{i,1}}} \in (0, 1),$$

where the limits are taken for $p_{i,j} \to 0^+$ for every $j \neq j_{i,1}, j_{i,2}$. Thus, for every $i = 1, \ldots, n$, there exist $\{p_{i,j}\}_{j=1}^{m_i}$ such that if we define a naive Bayes with class variable $C \in \{-1, +1\}$, $P(C = -1) = P(C = +1) = \frac{1}{2}$; predictors $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$, with $P(X_i = \xi_i^j | C = -1) = p_{i,j}$ and $P(X_i = \xi_i^j | C = +1) = e^{\alpha_i(j)} p_{i,j}$, we have that $f = sgn\left(\sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) \right) \right)$ is the decision function induced by the naive Bayes classifier defined above.

Proof by contradiction could be used to demonstrate that *ii)* implies *i)*. If $\sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} \alpha_i(j) l_j^{\Omega_i}(x_i) \right)$ is the polynomial that sign-represents f built as in Equation (7) and we assume $e^{\alpha_{i_0,j}} > 1$ for every $j \in \Omega_i$ and for a fixed i_0, we have $\alpha_{i_0,j} \geq 0$ for every j and, from the definition of $\alpha_i(j)$ (proof of Theorem 3), we get

$$P(X_{i_0} = \xi_i^j | C = +1) > P(X_{i_0} = \xi_i^j | C = -1).$$

Summing now over j, we obtain

$$1 = \sum_{j=1}^{m_i} P(X_{i_0} = \xi_i^j | C = +1) > \sum_{j=1}^{m_i} P(X_{i_0} = \xi_i^j | C = -1) = 1,$$

which is absurd, proving the statement.

■

3.3 Tree augmented naive Bayes

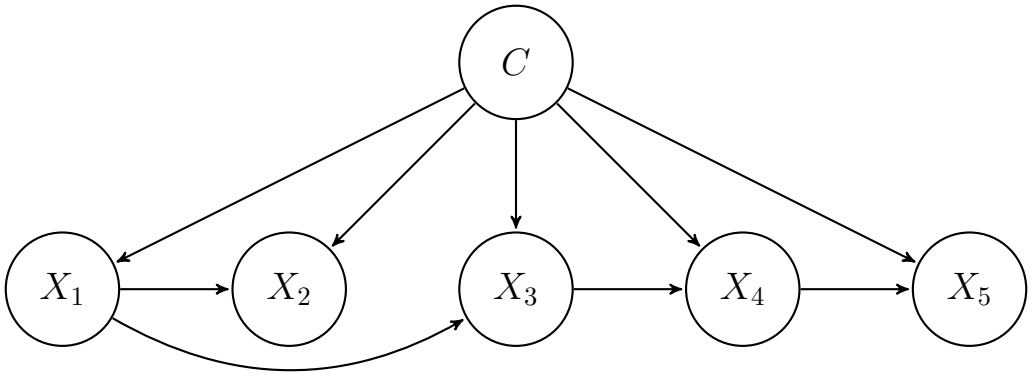


Figure 3: Tree augmented naive Bayes classifier structure with five predictor variables

We now consider a tree augmented naive Bayes (TAN) classifier (Friedman et al., 1997) as shown in Figure 3. In this model, a predictor variable $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$ is allowed to have at most two parents, the class C and an other variable, $X_{pa(i)} \in \Omega_{pa(i)}$. The joint probability distribution of $(C, X_1, X_2, \ldots, X_n)$ over $\{-1, +1\} \times \Omega_1 \times \cdots \times \Omega_n$ can be factorized according to the Bayesian network theory as

$$P(C = c) \prod_{i=1}^{n} P\left(X_i = x_i | C = c, X_{pa(i)} = x_{pa(i)}\right). \tag{9}$$

We can write down a similar representation to the NB case. For each $i = 1, \ldots, n$, we apply Equation (4) and obtain

$$P\left(X_i = x_i | C = c, X_{pa(i)} = x_{pa(i)}\right) = \prod_{j=1}^{m_i} \prod_{k=1}^{m_{pa(i)}} p_i(j|c,k)^{\left(\ell_k^{\Omega_{pa(i)}}(x_{pa(i)}) \ell_j^{\Omega_i}(x_i)\right)}. \tag{10}$$

We can now prove, combining Equations (9) and (10), a result similar to the NB case.

Lemma 5 *If f^{TAN} is the decision function induced by a TAN for a binary classification problem with n categorical predictor variables $\{X_i \in \Omega_i\}_{i=1}^n$, then there exists a polynomial, of the form*

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i) \sum_{k=1}^{m_{pa(i)}} \beta_i(j|k) \ell_k^{\Omega_{pa(i)}}(x_{pa(i)}),$$

that sign-represents f^{TAN}, where we consider $\sum_{k=1}^{m_{pa(i)}} \beta_i(j|k)\ell_k^{\Omega_{pa(i)}}(x_{pa(i)}) = \beta_i(j)$ when $\Omega_{pa(i)} = \emptyset$, that is, when class C is the only parent of a node (the root node of the tree).

Proof The proof is a straightforward computation of the logarithm of Equation (9) using Equation (10) and the definition $\beta_i(j|k) = \ln\left(\frac{p_i(j|+1,k)}{p_i(j|-1,k)}\right)$. ■

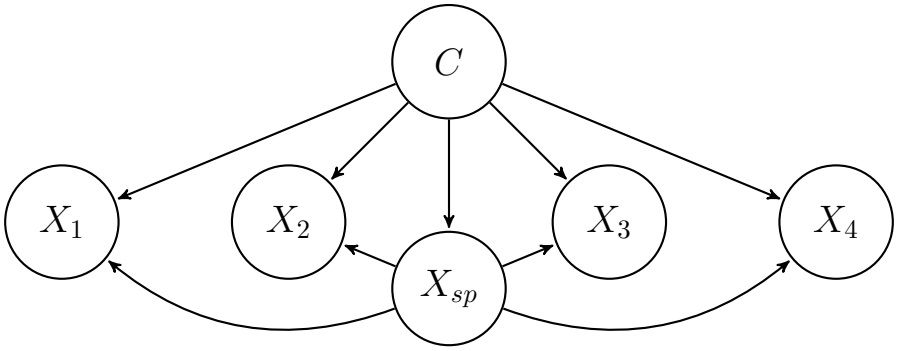


Figure 4: SPODE Bayes classifier structure with five predictor variables

A particular case of TAN is the *SuperParent-One-Dependence Estimator* (SPODE) (Keogh and Pazzani, 2002), where all the predictors depend on the same predictor (superparent) (Figure 4). The joint distribution factorizes as follows:

$$P\left(C=c\right)P\left(X_{sp}=x_{sp}|C=c\right)\prod_{i\neq sp}P\left(X_i=x_i|C=c,X_{sp}=x_{sp}\right),$$

where X_{sp} stands for the superparent node. In this case, the representation of Lemma 5 reduces to

$$f^{SPODE}(\mathbf{x}) = sgn\left(\sum_{i\neq sp}\sum_{j=1}^{m_i}\ell_j^{\Omega_i}(x_i)\sum_{k=1}^{m_{sp}}\beta_i(j|k)\ell_k^{\Omega_{sp}}(x_{sp})\right), \tag{11}$$

where f^{SPODE} is the induced decision function. If we fix the superparent node, we have a stronger characterization of the induced decision functions, the analogue of Theorem 3.

Theorem 6 *A decision function for a binary classification problem over categorical predictor variables is sign-represented by a polynomial of the form*

$$\sum_{i\neq sp}\sum_{j=1}^{m_i}\ell_j^{\Omega_i}(x_i)\sum_{k=1}^{m_{pa(i)}}\beta_i(j|k)\ell_k^{\Omega_{sp}}(x_{sp}),$$

if and only if it is induced by a SPODE classifier with X_{sp} as the superparent node.

Proof The *if* part of the theorem is precisely Equation (11). To prove the *only if* part we repeat a similar argument as in Theorem 3. We observe (Lemma 2, point 4, with $J = \{i\}$ and $I = \{i, sp\}$) that for every $i \neq sp$,

$$\ell_k^{\Omega_{sp}}(x_{sp}) = \sum_{j=1}^{m_i}\ell_j^{\Omega_i}(x_i)\ell_k^{\Omega_{sp}}(x_{sp}),$$

and so the coefficient $\beta_i(j|k)$ could be seen as

$$\beta_i(j|k) = \ln\left(\frac{P(X_i = j|X_{sp} = k, C = +1)}{P(X_i = j|X_{sp} = k, C = -1)}\right) + \alpha_i(k),$$

where $\sum_{i\neq sp}\alpha_i(k) = \ln\left(\frac{P(X_{sp}=\xi^k_{sp}|C=+1}{P(X_{sp}=\xi^k_{sp}|C=-1}\right) + \alpha$ and $\alpha = \ln\left(\frac{P(C=+1)}{P(C=-1)}\right)$. Then adjusting $\alpha_i(k)$ and α properly we can find a SPODE model, that is, probability distributions over the predictors and the class that induces

$$f = sgn\left(\sum_{i\neq sp}\sum_{j=1}^{m_i}\sum_{k=1}^{m_{sp}}\beta_i(j|k)\ell_k^{\Omega_{sp}}(x_{sp})\ell_j^{\Omega_i}(x_i)\right),$$

for every $\beta_i(j|k) \in \mathbb{R}$. ■

3.4 Bayesian Network-Augmented Naive Bayes

If the predictor sub-graph can be a generic Bayesian network, we have a Bayesian network-augmented naive Bayes (BAN) classifier. In this case the joint probability distribution is factorized as follows:

$$P(C = c)\prod_{i=1}^{n} P\left(X_i = x_i|C = c, \mathbf{X}_{\mathbf{pa}(i)} = \mathbf{x}_{\mathbf{pa}(i)}\right), \tag{12}$$

where $\mathbf{X}_{\mathbf{pa}(i)}$ denotes the vector of the parent variables of X_i that are not C. From now on we will write $\mathbf{pa}(i)$ for the set of indexes defining X_i's parents that are not C and $\mathbb{M}_i = \times_{s\in\mathbf{pa}(i)}\{1,\ldots,m_s\}$ for the set of possible configurations of the parents of X_i. Applying the same arguments as in previous sections we can prove the lemma below.

Lemma 7 *If f^{BAN} is the decision function induced by a BAN classifier for a classification problem with n categorical predictors variables $\{X_i \in \Omega_i \subset \mathbb{R}, |\Omega_i| = m_i\}_{i=1}^n$, then there exists a polynomial of the form*

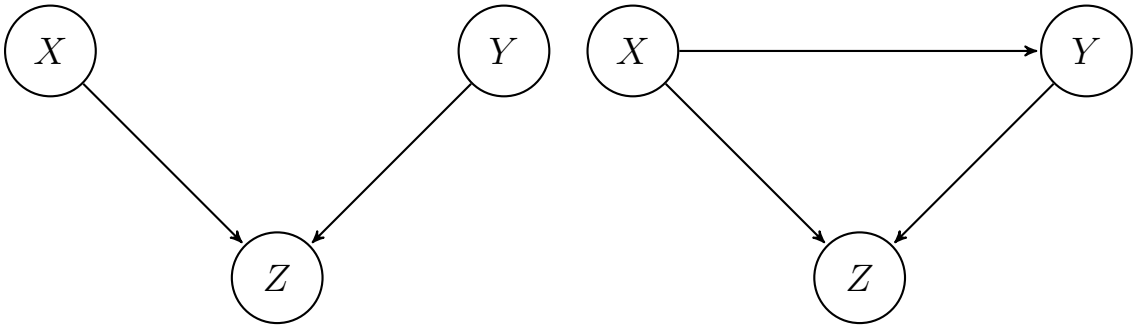
$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\ell_j^{\Omega_i}(x_i)\sum_{\mathbf{k}\in\mathbb{M}_i}\beta_i(j|\mathbf{k})\prod_{s\in\mathbf{pa}(i)}\ell_{k_s}^{\Omega_s}(x_s),$$

which sign-represents f^{BAN}, where we write $\sum_{\mathbf{k}\in\mathbb{M}_i}\beta_i(j|\mathbf{k})\prod_{s\in\mathbf{pa}(i)}\ell_{k_s}^{\Omega_s}(x_s) = \beta_i(j)$ when a variable does not have parents that are not C, that is, $\mathbf{pa}(i) = \emptyset$.

Proof Given a BAN model over predictors $X_i \in \Omega_i = \{\xi_i^1,\ldots,\xi_i^{m_i}\}$, we define

$$\beta_i(j|\mathbf{k}) = \ln\left(\frac{P\left(X_i = \xi_i^j|C = +1, X_s = \xi_s^{k_s}, \forall s \in \mathbf{pa}(i)\right)}{P\left(X_i = \xi_i^j|C = -1, X_s = \xi_s^{k_s}, \forall s \in \mathbf{pa}(i)\right)}\right).$$

Using Equation (4) and taking the logarithm of Equation (12) we obtain the polynomial representation. ■



(a) V-structure (b) Not a V-structure

Generally speaking, it is not always possible to prove results similar to Theorem 3 or Theorem 6 for BAN classifiers, when decision functions are completely characterized by the set of sign-representing polynomials. Like Yang and Wu (2012), we find that problems arise in the presence of V-structures (Figure 5a) in the predictor sub-graph. A V-structure appears when two nodes share the same child, but are not directly connected. In absence of V-structures we can prove the following result, which extend the previous ones,

Theorem 8 *Let $\mathcal{G}$ be a directed acyclic graph with node X_i for $i \in \{1, 2, \ldots, n\}$, and let f be a decision function over predictor variables $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$. Suppose that $\mathcal{G}$ does not contain V-structures, then we have that f is sign-represented by the following polynomial*

$$r(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k} \in \mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s),$$

if and only if f is induced by a BAN classifier whose predictor sub-graph is $\mathcal{G}$.

Proof We merely have to prove the *only if* because the *if* implication is precisely Lemma 7. Given a polynomial of the form

$$r(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j \in \Omega_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k} \in \mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s),$$

we have to find a BAN classifier inducing $sgn(r(\mathbf{x}))$, whose predictor sub-graph is $\mathcal{G}$. We merely have to define the conditional probability distribution of every variable given its parents, since the structure of the BAN is already fixed by $\mathcal{G}$. For every $i \in \{1, 2, \ldots, n\}$, we observe that the sub-graph of the parents of X_i is a fully connected Bayesian network, otherwise we will have a V-structure on $\mathcal{G}$. For every i, we can rewrite using Lemma 2 the ith addend on the summation,

$$\begin{aligned}
&\sum_{j \in \Omega_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k} \in \mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s \in \mathbf{pa(i)}} \ell_{k_s}^{\Omega_s}(x_s) + \sum_{\mathbf{k} \in \mathbb{M}_i} \alpha_i(\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s) - \sum_{\mathbf{k} \in \mathbb{M}_i} \alpha_i(\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s) \\
&= \sum_{j \in \Omega_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k} \in \mathbb{M}_i} (\beta_i(j|\mathbf{k}) + \alpha_i(\mathbf{k})) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s) - \sum_{\mathbf{k} \in \mathbb{M}_i} \alpha_i(\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s).
\end{aligned}$$

Using the *free parameters* $\alpha_i(\mathbf{k})$, it is possible to find for every $\mathbf{k}$, $p_i(j|\mathbf{k},+1)$ and $p_i(j|\mathbf{k},-1) \in (0,1)$ such that

$$\sum_{j=1}^{m_i} p_i(j|\mathbf{k},+1) = \sum_{j=1}^{m_i} p_i(j|\mathbf{k},-1) = 1$$

$$\beta_i(j|\mathbf{k}) + \alpha_i(\mathbf{k}) = \ln \frac{p_i(j|\mathbf{k},+1)}{p_i(j|\mathbf{k},-1)}.$$

To avoid changing the polynomial $r(\mathbf{x})$, we have to subtract

$$\sum_{\mathbf{k}\in\mathbb{M}_i} \alpha_i(\mathbf{k}) \prod_{s\in\mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s)$$

from another addend on the summation. Because the parents of X_i are fully connected, we have that among the other addends of $r(\mathbf{x})$, apart from the ith, there is one product that contains $\prod_{s\in\mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s)$ and so we just subtract $\alpha_i(\mathbf{k})$ from the related coefficient. Iterating the above procedure for all the nodes of the graph $\mathcal{G}$, we are able to build a probability distribution over $X_1, X_2, \ldots, X_n, C$ that satisfies the Bayesian network structure given by $\mathcal{G}$. More precisely, setting

$$P\left(X_i = \xi_i^j | C = c, X_s = \xi_s^{k_s}, \forall s \in \mathbf{pa}(i)\right) = p_i(j|\mathbf{k}, c),$$

we obtain the target BAN model. ■

3.5 Full Bayesian Network

When the predictor sub-graph is a fully connected Bayesian network (Figure 6), that is, a directed acyclic graph with the maximum number of arcs, we can apply Theorem 8 and obtain a representation of polynomials sign-representing the induced decision function. Observing that a fully connected Bayesian network classifier (FBN) is in fact a general classifier, that is, a classifier able to induce any decision function over $\Omega = \times_{i=1}^n \Omega_i$ whatsoever, we have the following corollary.

Corollary 9 *If Φ is a classifier for a binary class problem with n predictor variables $X_1, \ldots, X_n$ such that $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\} \subset \mathbb{R}$, $|\Omega_i| = m_i$, then the associated decision function, f^Φ, is sign-represented by a polynomial of the form*

$$\sum_{\mathbf{k}\in\mathbb{M}} \gamma_{\mathbf{k}} \prod_{i=1}^n \ell_{k_i}^{\Omega_i}(x_i),$$

where $\mathbb{M} = \times_{i=1}^n \{1, \ldots, m_i\}$.

Since the product of the Lagrange bases, $\prod_{i=1}^n \ell_{k_i}^{\Omega_i}(x_i)$, interpolates the Iverson bracket over all the predictors, that is,

$$\prod_{i=1}^n \ell_{k_i}^{\Omega_i}(x_i) = [x_i = \xi_i^{k_i}, \forall i = 1, \ldots, n],$$

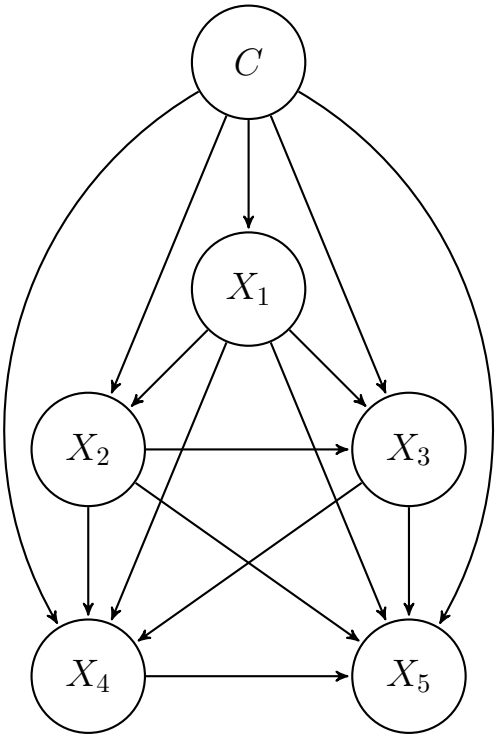


Figure 6: FBN classifier structure with five predictor variables

we have that the coefficients $\gamma_{\mathbf{k}}$ in Corollary 9 are the values of the polynomial at point $(\xi_1^{k_1}, \xi_2^{k_2}, \ldots, \xi_n^{k_n})$, and so $f^{\Phi}(\xi_1^{k_1}, \xi_2^{k_2}, \ldots, \xi_n^{k_n}) = sgn(\gamma_{\mathbf{k}})$. Roughly speaking, a new instance $(\xi_1^{k_1}, \xi_2^{k_2}, \ldots, \xi_n^{k_n})$ will be classified as $C = +1$ if and only if $\gamma_{\mathbf{k}} > 0$. Moreover the set

$$\mathcal{P}^{FBN} = \left\{ \sum_{\mathbf{k} \in \Omega} \gamma_{\mathbf{k}} \prod_{i=1}^{n} \ell_{k_i}^{\Omega_i}(x_i) \text{ s.t. } \gamma_{\mathbf{k}} \in \mathbb{R} \right\}$$

of polynomials, which could sign-represent every classifier, is a space of dimension $M = \prod_{i=1}^{n} m_i$. From now on we will write

$$\delta_{\mathbf{k}}(\mathbf{x}) = \prod_{i=1}^{n} \ell_{k_i}^{\Omega_i}(x_i), \tag{13}$$

for the $\mathbf{k}$-th element of the canonical basis of $\mathcal{P}^{FBN}$. We call $\{\delta_{\mathbf{k}}\}_{\mathbf{k} \in \Omega}$ the canonical basis because the sign of the coefficients with respect to this basis is the value of the sign-represented decision function. Corollary 9 states that $sgn(\mathcal{P}^{FBN}) = \{-1, 1\}^{\Omega}$.

4. Expressive Power of Bayesian Network Classifiers

So far, we have seen how to build polynomial threshold functions that sign-represent decision functions induced by Bayesian network classifiers. We use now the resulting representation to bound the number of decision functions representable by Bayesian network classifiers.

As observed, Corollary 9 says that $sgn(\mathcal{P}^{FBN}) = \{-1,1\}^{\Omega}$. We now study NB, SPODE and BAN through the families of associated polynomial threshold functions. Moreover, we embed those families in $\mathcal{P}^{FBN}$. For predictor variables $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$, $i = 1, \ldots, n$, for every $sp \in \{1, \ldots, n\}$ and a directed acyclic graph $\mathcal{G}$ without V-structures we define

$$\mathcal{P}^{NB} = \left\{ r(\mathbf{x}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) \right) \text{ s.t. } \alpha_i(j) \in \mathbb{R} \right\}, \tag{14}$$

$$\mathcal{P}_{sp}^{SPODE} = \left\{ r(\mathbf{x}) = \sum_{i \neq sp} \sum_{j=1}^{m_i} \sum_{k=1}^{m_{sp}} \beta_i(j|k) \ell_k^{\Omega_{sp}}(x_{sp}) \ell_j^{\Omega_i}(x_i) \text{ s.t. } \beta_i(j|k) \in \mathbb{R} \right\}, \tag{15}$$

$$\mathcal{P}_{\mathcal{G}}^{BAN} = \left\{ r(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k} \in \mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s) \text{ s.t. } \beta_i(j|\mathbf{k}) \in \mathbb{R} \right\}, \tag{16}$$

where $\mathbf{pa}(i)$ is a function that maps every i into the set of parents of X_i in the directed acyclic graph $\mathcal{G}$, and $\mathbb{M}_i = \times_{s \in \mathbf{pa}(i)} \{1, \ldots, m_s\}$. The families $\mathcal{P}^{NB}$, $\mathcal{P}_{sp}^{SPODE}$ and $\mathcal{P}_{\mathcal{G}}^{BAN}$ are the sets of polynomials sign-representing the decision functions induced by naive Bayes classifier, SPODE classifier and BAN classifier, respectively. Hence $sgn(\mathcal{P}^{NB})$, $sgn(\mathcal{P}_{sp}^{SPODE})$ and $sgn(\mathcal{P}_{\mathcal{G}}^{BAN})$ are the sets of decision functions induced by naive Bayes, SPODE and BAN classifiers, respectively. Obviously, we have that

$$\mathcal{P}^{NB} \subset \mathcal{P}_{sp}^{SPODE} \subset \mathcal{P}_{\mathcal{G}}^{BAN} \subset \mathcal{P}^{FBN},$$

and

$$sgn(\mathcal{P}^{NB}) \subset sgn(\mathcal{P}_{sp}^{SPODE}) \subset sgn(\mathcal{P}_{\mathcal{G}}^{BAN}) \subset sgn(\mathcal{P}^{FBN}) = \{-1,+1\}^{\Omega}.$$

We can prove that the above sets are indeed subspaces of $\mathcal{P}^{FBN}$ and we can compute their dimensions.

Lemma 10 *$\mathcal{P}^{NB}$ is a subspace of $\mathcal{P}^{FBN}$ of dimension $\sum_{i=1}^{n} m_i - n + 1$.*

Proof Obviously $\mathcal{P}^{NB} = \left\{ p(\mathbf{x}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) \right), \alpha_i(j) \in \mathbb{R} \right\}$ is a subspace of $\mathcal{P}^{FBN}$. The union of the Lagrange bases over different variables is not a basis, because for each $i = 1, \ldots, n$ we have that

$$1 = \sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i) \text{ for every } x_i \in \mathbb{R}.$$

So for every i, we can define

$$\mathcal{B}_i = \left\{ \bigcup_{j=2}^{m_i} \{l_j^{\Omega_i}(x_i)\} \right\} \cup \{e_0\},$$

where e_0 is the polynomial constant 1, and we find that $\mathcal{B}_i$ is a basis of polynomials in x_i of degree $|\Omega_i| - 1 = m_i - 1$, equivalent to the Lagrange basis over Ω_i. Then, we have that

$$\mathcal{B} = \bigcup_{i=1}^{n} \mathcal{B}_i = \bigcup_{i=1}^{n} \bigcup_{j=2}^{m_i} \left\{ l_j^{\Omega_i}(x_i) \right\} \cup \{e_0\}$$

generates the subspace $\mathcal{P}^{NB}$. We prove that $\mathcal{B}$ is in fact a basis of $\mathcal{P}^{NB}$. We have to prove that the elements of $\mathcal{B}$ are linearly independent. We consider

$$p(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \sum_{j=2}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) + \alpha_0 e_0 = 0, \, \forall (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n.$$

If, as usual, $\Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$, let us consider $p(x_1, \ldots, x_n)$ evaluated in $(\xi_1^1, \xi_2^1, \ldots, \xi_n^1)$,

$$0 = p(\xi_1^1, \xi_2^1, \ldots, \xi_n^1) = \sum_{i=1}^{n} \sum_{j=2}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(\xi_i^1) + \alpha_0 e_0 = \alpha_0,$$

since $\ell_j^{\Omega_i}(\xi_i^1) = 0$ for every $j \neq 1$. And so $\alpha_0 = 0$. We now evaluate $p(\cdot)$ over $(\xi_1^j, \xi_2^1, \ldots, \xi_n^1)$ and we have that, for every $j = 2, \ldots, m_i$,

$$0 = p(\xi_1^j, \xi_2^1, \ldots, \xi_n^1) = \alpha_1(j),$$

since $\ell_j^{\Omega_1}(\xi_1^j) = 1$ for every $j = 2, \ldots, m_1$. We repeat the above argument for every variable x_i, $i = 1, \ldots, n$ and we obtain $\alpha_i(j) = 0$ for every $i = 1, \ldots, n$ and every $j = 2, \ldots, m_i$. We have proved that the elements of $\mathcal{B}$ generate $\mathcal{P}^{NB}$ and are linearly independent, so they form a basis of $\mathcal{P}^{NB}$. Consequently we obtain

$$\dim(\mathcal{P}^{NB}) = |\mathcal{B}| = \sum_{i=1}^{n} m_i - n + 1.$$

■

Analogously we can prove, in the general case,

Lemma 11 *For every Bayesian network classifier without V-structures in the predictor sub-graph $\mathcal{G}$, the set $\mathcal{P}_{\mathcal{G}}^{BAN}$ is a subspace of $\mathcal{P}^{FBN}$ of dimension*

$$\sum_{i=1}^{n} \left((m_i - 1) \prod_{s \in \mathbf{pa}(i)} m_s \right) + 1.$$

And, in the particular case of $SPODE$, we have,

Lemma 12 *For every $sp = 1, \ldots, n$, the set $\mathcal{P}_{sp}^{SPODE}$ is a subspace of $\mathcal{P}^{FBN}$ of dimension $m_{sp} \left(1 - n + \sum_{i \neq sp} m_i\right)$.*

We now consider the space $\mathcal{P}^{FBN}$ with respect to the canonical basis given by Equation (13). With respect to this coordinate system we have that each orthant represents a decision function. We know that the number of orthants of a M-dimensional space is 2^M, the number of decision functions over a set of cardinality M. Since we now have a bijection between orthants in $\mathcal{P}^{FBN}$ and decision functions over Ω, in order to compute how many decision functions are representable by a class of Bayesian network classifier (NB, SPODE or BAN) we merely have to count the number of orthants in $\mathcal{P}^{FBN}$ intersected by the corresponding subspaces ($\mathcal{P}^{NB}$, $\mathcal{P}^{SPODE}_{sp}$, $\mathcal{P}^{BAN}_{\mathcal{G}}$).

Theorem 13 (Flatto (1970)) *A d-dimensional subspace in an M-space intersects at most $C(M,d) = 2\sum_{k=0}^{d-1}\binom{M-1}{k}$ orthants with equality if and only if it is in general position.*

Definition 14 *A d-dimensional subspace V of $\mathbb{R}^M$ is in general position if the M subspaces $V \cap H_i$, where $H_i = \{\mathbf{x} \in \mathbb{R}^n \text{ s.t. } x_i = 0\}$ are hyperplanes of V in general position, that is, all the intersections of d of such hyperplanes are the zero vector.*

Applying Theorem 13 to our case, we find that the space $\mathcal{P}^{FBN}$ is minimal in the following sense.

Corollary 15 *If V is a d-dimensional subspace of $\mathcal{P}^{FBN}$, then $|sgn(V)| \leq C(M,d)$, where $M = dim(\mathcal{P}^{FBN})$ and equality holds if and only if V is in general position with respect to the canonical basis of $\mathcal{P}^{FBN}$.*

As a first result of Corollary 15 we have that the space $\mathcal{P}^{FBN}$ is the *smallest* vectorial space of polynomials in $x_1, \ldots, x_n$ that sign-represents every decision function over Ω, that is, there is not a space V of polynomials in $x_1, \ldots, x_n$ with degrees in each variable x_i that is less or equal than $m_i - 1$ such that $sgn(V) = \{-1, +1\}^\Omega$ and $dim(V) < dim(\mathcal{P}^{FBN})$. This justifies the choice of $\mathcal{P}^{FBN}$ as the space to study the polynomial families defined in Equations (14), (15) and (16).

Next, we can use Corollary 15 combined with Lemma 11 to upper bound the number of decision functions that are sign-representable by BAN classifiers with a fixed predictor sub-graph $\mathcal{G}$ not containing V-structures.

Corollary 16 *Consider a BAN classifier over predictor variables $X_i \in \Omega_i$, $|\Omega_i| = m_i$ for every $i = 1, \ldots, n$. Moreover suppose that the predictor sub-graph $\mathcal{G}$ does not contain V-structures. Then we have*

$$2^d \leq |sgn(\mathcal{P}^{BAN}_{\mathcal{G}})| \leq C(M,d) = 2\sum_{k=0}^{d-1}\binom{M-1}{k},$$

where $d = \sum_{i=1}^{n}\left((m_i - 1)\prod_{s\in\mathbf{pa}(i)} m_s\right) + 1$ and $M = \prod_{i=1}^{n} m_i$.

Corollary 16 extends an observation of Peot (1996) about the fraction of decision functions representable by Bayesian network classifiers as follows.

Corollary 17 *We consider, for every $n \in \mathbb{N}$, classification problems with predictors $X_i \in \Omega_i \subset \mathbb{R}$, $|\Omega_i| = m_i$ for $i = 1, \ldots, n$. For every n, let $\mathcal{G}_n$ be a directed acyclic graph over the predictor variables, not containing V-structures. Suppose moreover that if $\mathbf{pa}_n(i)$ are the functions that map every X_i into the set of parents in the graph $\mathcal{G}_n$,*

$$|\mathbf{pa}_n(i)| \leq K \quad \forall n \in \mathbb{N} \text{ and } i \in \{1, \ldots, n\},$$

then we have that

$$\lim_{n\to\infty} \frac{|sgn(P^{BAN}_{\mathcal{G}_n})|}{|\{-1,+1\}^{\Omega(n)}|} = \lim_{n\to\infty} \frac{|sgn(P^{BAN}_{\mathcal{G}_n})|}{2^{|\Omega(n)|}} = 0,$$

where $\Omega(n) = \times_{i=1}^{n} \Omega_i$. In other words, the fraction of decision functions representable by BAN classifiers, with a fixed maximum number of parents for each variable, becomes vanishingly small by increasing the number of predictors.

Proof For every $n \in \mathbb{N}$, we apply Corollary 16 and we obtain

$$\left|sgn(\mathcal{P}^{BAN}_{\mathcal{G}_n})\right| \leq C\left(M(n), d(n)\right) = 2 \sum_{k=0}^{d(n)-1} \binom{M(n)-1}{k},$$

where $d(n) = \sum_{i=1}^{n} \left((m_i - 1) \prod_{s\in\mathbf{pa}(i)} m_s\right) + 1$ and $M(n) = |\Omega(n)| = \prod_{i=1}^{n} m_i$. We observe now that, as $n \to \infty$,

$$\frac{d(n)}{M(n)} \to 0$$

and thus,

$$\frac{C(M(n), d(n))}{2^{M(n)}} \to 0,$$

which proves the statement. ■

4.1 VC dimension

In this section we compare our results, especially the upper bound in Corollary 16, with the classical method for evaluating the expressive power of classifiers, that is, the Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1971). Firstly we need some definitions.

Definition 18 *Given a subset of decision functions $\mathcal{F} \subset \{-1,+1\}^{\Omega}$, we say that $\mathcal{F}$ shatters $\Omega_0 \subset \Omega$ if for every $g \in \{-1,+1\}^{\Omega_0}$ there exists a decision function $f \in \mathcal{F}$ such that $f_{|\Omega_0} = g$.*

Roughly speaking, $\mathcal{F} \subset \{-1,+1\}^{\Omega}$ shatters a subset $\Omega_0 \subset \Omega$ if the decision functions in $\mathcal{F}$ could recognize every possible classification of Ω_0.

The VC dimension of a set of decision functions is defined as the cardinality of the largest subset of Ω shattered.

Definition 19 *The VC dimension of $\mathcal{F} \subset \{-1,+1\}^{\Omega}$, denoted by $d_{VC}(\mathcal{F})$, is defined by*

$$d_{VC}(\mathcal{F}) = \max\{|\Omega_0| \text{ s.t. } \Omega_0 \text{ is shattered by } \mathcal{F}\}.$$

VC dimension is useful in machine learning theory. It is a fundamental concept of Vapnik-Chervonenkis theory (Vapnik and Chervonenkis, 1971). The following result, know as the Sauer-Shelah-Vapnik-Chervonenkis lemma (Sauer, 1972) bounds the cardinality of a family of decision functions with a given VC dimension.

Theorem 20 (Sauer-Shelah-Vapnik-Chervonenkis) *If $\mathcal{F} \subset \{-1,+1\}^{\Omega}$, $|\Omega| = M$ and $d_{VC}(\mathcal{F}) = d$ then*

$$|\mathcal{F}| \leq \sum_{i=0}^{d} \binom{M}{i}.$$

In our case we have that the sets $sgn(\mathcal{P}^{NB})$, $sgn(\mathcal{P}^{SPODE}_{sp})$ and $sgn(\mathcal{P}^{BAN}_{\mathcal{G}})$ are generated by linear subspaces of given dimensions. In this case, we can prove that the VC dimension and the geometric dimension are the same.

Lemma 21 *If V is a d-dimensional space of $\mathcal{P}^{FBN}$, which does not lie in any hyperplane $H_{\mathbf{k}'} = \left\{p = \sum_{\mathbf{k}\in\mathbb{M}} \gamma_{\mathbf{k}}\delta_{\mathbf{k}}(\mathbf{x}) \in \mathcal{P}^{FBN} \text{ s.t. } \gamma_{\mathbf{k}'} = 0\right\}$ for $\mathbf{k}' \in \mathbb{M}$, we have that*

$$d_{VC}(sgn(V)) = d.$$

Proof Since V does not lie in any $H_{\mathbf{k}'}$, that is,

$$V \cap \left\{ p = \sum_{\mathbf{k}\in\mathbb{M}} \gamma_{\mathbf{k}}\delta_{\mathbf{k}}(\mathbf{x}) \in \mathcal{P}^{FBN} \text{ s.t. } \gamma_{\mathbf{k}'} = 0 \right\} \neq V,$$

we have that there is at least one subset $\Omega' \subset \Omega$ of cardinality d that is shattered by $sgn(V)$ and thus

$$d_{VC}(sgn(V)) \geq d.$$

We prove now that $d_{VC}(sgn(V)) \leq d$. We consider $\Omega_0 \subset \Omega$ of cardinality $d+1$. Remember that considering the canonical basis over $\mathcal{P}^{FBN}$ we have that each coefficient $\gamma_{\mathbf{k}}$ is the value of the polynomial $r(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{M}} \gamma_{\mathbf{k}}\delta_{\mathbf{k}}(\mathbf{x})$ at point $(\xi_1^{k_1}, \ldots, \xi_n^{k_n})$. Because V is a subspace of dimension d, for every $r(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{M}} \gamma_{\mathbf{k}}\delta_{\mathbf{k}}(\mathbf{x}) \in V$, we have that the $d+1$ coefficients of r related to the points in Ω_0 are linearly dependent, that is, there is $(\xi_1^{j_1}, \ldots, \xi_n^{j_n}) \in \Omega_0$ such that

$$\gamma_{(j_1,\ldots,j_n)} = \sum_{\mathbf{k}\in\mathbb{M}_0\setminus(j_1,\ldots,j_n)} \alpha_{\mathbf{k}}\gamma_{\mathbf{k}}, \tag{17}$$

where $\mathbb{M}_0 = \left\{(k_1,\ldots,k_n) \text{ s.t. } (\xi_1^{k_1}, \ldots, \xi_n^{k_n}) \in \Omega_0\right\}$. Consider now $g : \Omega_0 \to \{-1,+1\}$, such that

$$g(\mathbf{k}) = \begin{cases} sgn(\alpha_{\mathbf{k}}) & \mathbf{k} \in \mathbb{M}_0 \setminus (j_1,\ldots,j_n) \\ -1 & \mathbf{k} = (j_1,\ldots,j_n). \end{cases}$$

Suppose that there exists $f \in sgn(V)$ such that $f_{|\Omega_0} = g$, then there exists $r(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{M}} \gamma_{\mathbf{k}}\delta_{\mathbf{k}}(\mathbf{x}) \in V$ such that $g(\mathbf{x}) = f(\mathbf{x}) = sgn(r(\mathbf{x}))$ for every $\mathbf{x} \in \Omega_0$ and so

$$sgn(\gamma_{\mathbf{k}}) = sgn(\alpha_{\mathbf{k}}) \text{ for every } \mathbf{k} \in \mathbb{M}_0 \setminus (j_1, \ldots, j_n)$$

$$sgn(\gamma_{(j_1,\ldots,j_n)}) = -1.$$

But from Equation (17) we have that

$$sgn(\gamma_{(j_1,\ldots,j_n)}) = sgn\left(\sum_{\mathbf{k}\in\mathbb{M}_0\setminus(j_1,\ldots,j_n)} \alpha_{\mathbf{k}}\gamma_{\mathbf{k}}\right) = +1.$$

And so $sgn(V)$ does not shatter Ω_0.

■

Remark 22 *We observe that the condition required for V in Lemma 21 is not very restrictive. If V lies in some $H_{\mathbf{k}'}$ for $\mathbf{k}' \in \mathbb{M}$, we have that, for every $p(\mathbf{x}) \in V$, there exists a point $(\xi_1^{k'_1}, \ldots, \xi_n^{k'_n}) \in \Omega$ (the $\mathbf{k}'$th point) such that*

$$p(\xi_1^{k'_1}, \ldots, \xi_n^{k'_n}) = 0,$$

and thus, $sgn(V) = \emptyset$ because of our definition of decision functions, that is, $f : \Omega \to \{-1, +1\}$. Consequently all the subspaces that sign-represent sets of decision functions will automatically satisfy that condition.

Combining Theorem 20 and Lemma 21 we can bound the number of decision functions representable by Bayesian network classifiers without V-structures as

$$|sgn(\mathcal{P}_{\mathcal{G}}^{BAN})| \leq \sum_{k=0}^{d} \binom{M}{k},$$

where $d = \sum_{i=1}^{n} \left((m_i - 1)\prod_{s\in\mathbf{pa}(i)} m_s\right) + 1$ and $M = \prod_{i=1}^{n} m_i$. We observe that the resulting bounds are worse than those obtained in Corollary 16. In fact, we have that

$$\sum_{k=0}^{d} \binom{M}{k} = 2\sum_{k=0}^{d-1} \binom{M-1}{k} + \binom{M-1}{d} \geq 2\sum_{k=0}^{d-1} \binom{M-1}{k}.$$

We have proved that if, as in the case we have studied, the sets of decision functions considered are generated by vectorial spaces,Theorem 13 improves the bounds provided by VC theory.

5. Conclusions

In this paper we have shown how to build polynomial threshold functions related to Bayesian network classifiers. Our results reveal connections between the algebraic structure of the decision functions induced by BN classifiers and the topology of the structure of the predictor sub-graph. In absence of V-structures in the predictor sub-graph we have also proved that the specific polynomial representation fully characterized the type of Bayesian network classifier. By representing classifiers by polynomial threshold functions, we can obtain bounds on the number of decision functions which can be induced by Bayesian network classifiers with a given structure. The resulting bounds are shown to be sharper than those obtained in VC theory. The bounding does not hold in presence of V-structures in the predictor sub-graph. Strong characterizations of induced decision functions cannot be proven due to the conditional independence of V-structure. Moreover we observe that the obtained polynomial representation permits to easily prove the results of Ling and Zhang (2003).

The bounds points to the fact, already conjectured by Peot (1996) for naive Bayes, that if we fix the maximum number of parents in a Bayesian network classifier, the type of classifier considered is not *scalable*, in other words, more complex classifiers are expected to perform better when dealing with a large number of predictor variables.

Moreover, the resulting bounds for the number of decision functions representable are strictly upper bounds since the subspaces generated by the different Bayesian networks considered are not in general position. What happens in the case of subspaces not in general position? Clearly we have to define some other property to characterize the *position* of a subspace with respect to orthants in some given basis and try to count the number of such intersected orthants. With similar geometric results we will be able to precisely count the number of decision functions representable by a given Bayesian network classifier, and we will be able to compute the gain in expressibility from simple to more complicated Bayesian network classifiers.

Acknowledgments

This research has been partially supported by the Spanish Ministry of Economy and Competitiveness through Cajal Blue Brain (C080020-09) and TIN2010-20900-C04-04 projects.

References

Concha Bielza and Pedro Larrañaga. Discrete Bayesian network classifier: A survey. *ACM Computing Surveys*, 2014. (in press).

Pedro Domingos and Michael Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

Leopold Flatto. A new proof of the transposition theorem. *Proceedings of the American Mathematical Society*, 24(1):29–31, 1970.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

Kenneth E. Iverson. *A Programming Language*. John Wiley & Sons, Inc., New York, 1962.

Manfred Jaeger. Probabilistic classifiers and the concepts they recognize. In Tom Fawcett and Nina Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning (ICML-03)*, pages 266–273. AAAI Press, 2003.

Eamonn J. Keogh and Michael J. Pazzani. Learning the structure of augmented Bayesian classifiers. *International Journal on Artificial Intelligence Tools*, 11(04):587–601, 2002.

Charles X. Ling and Huajie Zhang. The representational power of discrete Bayesian networks. *Journal of Machine Learning Research*, 3:709–721, 2003.

Marvin Minsky. Steps toward artificial intelligence. In *Computers and Thought*, pages 406–450. McGraw-Hill, 1961.

Atsuyoshi Nakamura, Michael Schmitt, Niels Schmitt, and Hans Ulrich Simon. Inner product spaces for Bayesian networks. *Journal of Machine Learning Research*, 6:1383–1403, 2005.

Ryan O'Donnell and Rocco A. Servedio. New degree bounds for polynomial threshold functions. *Combinatorica*, 30(3):327–358, 2010.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

Mark A. Peot. Geometric implications of the naive Bayes assumption. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, pages 414–419, San Francisco, 1996. Morgan Kaufmann Publishers Inc.

Giovanni Pistone, Eva Riccomagno, and Henry P. Wynn. Gröbner bases and factorisation in discrete probability and Bayes. *Statistics and Computing*, 11(1):37–46, 2001.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

Vladimir N. Vapnik and Alexy Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16 (2):264–280, 1971.

Chi Wang and A.C. Williams. The threshold order of a Boolean function. *Discrete Applied Mathematics*, 31(1):51–69, 1991.

Youlong Yang and Yan Wu. On the properties of concept classes induced by multivalued Bayesian networks. *Information Sciences: An International Journal*, 184(1):155–165, 2012.